How is a rational (big) data deployment approach like optimizing the generation mix of a power company?

SILICON VALLEY DATA SCIENCE

John Akred &
Stephen O'Sullivan
@SVDataScience

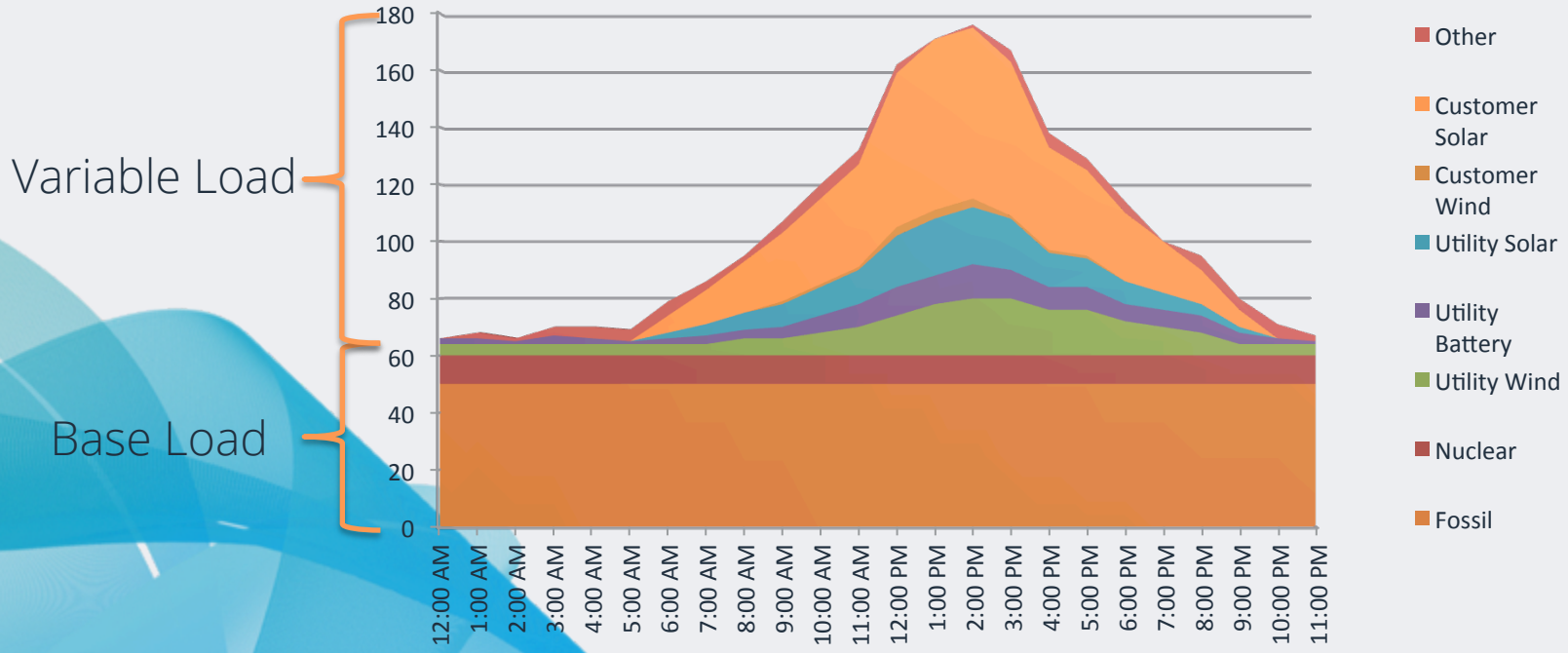
John Akred
@BigDataAnalysis

- Puppy Daddy
- PIF Husband
- Insider Trading (Investigation)
- VIX Index
- SPSS Clementine & Text Analysis for Surveys
- Condition Monitoring
- HR Analytics
- Smart Grid

Stephen Osullivan @steveos

- Father of 2
- Husband
- Database Engineer
- DBA
- Geek
- Quartermaster “Q”

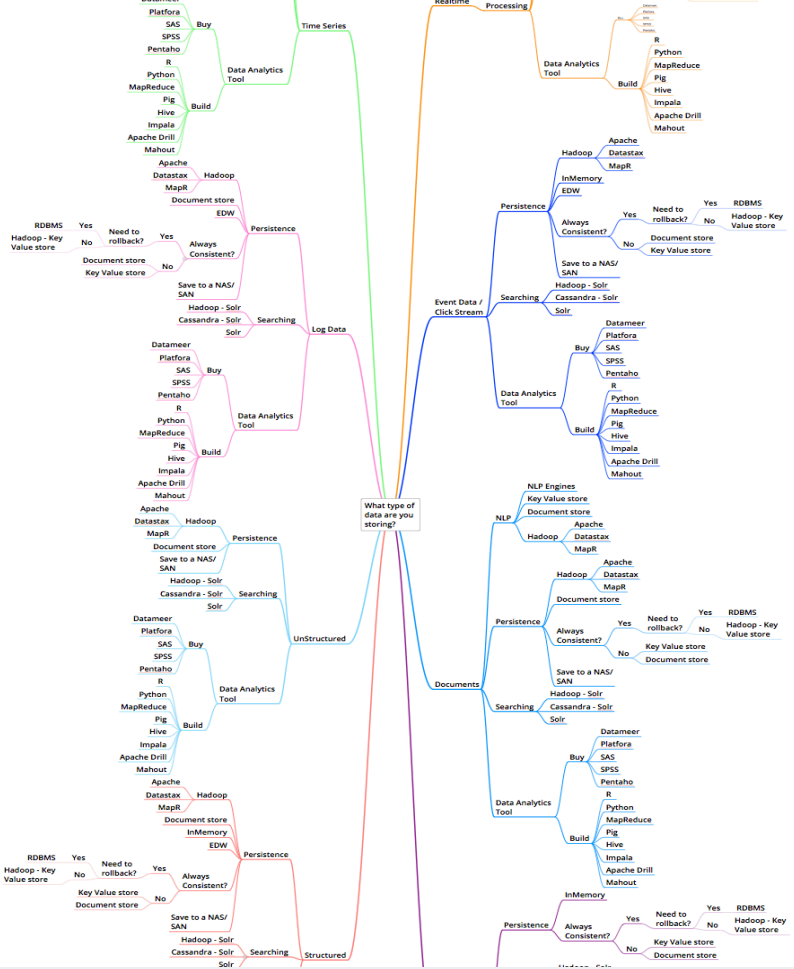
Utility Source Mix



PREVIOUSLY WE ASKED:

- What is the data type?
- What is the size of the data?
- What are the indexes?
- What are the foreign key constraints?

NOW WE ASK:



TOTAL COST OF OWNERSHIP

accenture
Technology Labs

Bare Metal vs. Cloud Smackdown

amazon web services™

This result debunks the idea that the cloud is not suitable for Hadoop MapReduce workloads given their heavy I/O requirements. Hadoop-as-a-Service provides a better price-performance ratio than the bare-metal counterpart.

http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Hadoop-Deployment-Comparison-Study.pdf

TUNING MATTERS!

We consistently observed that the performance improvement by applying various tuning techniques made a huge impact.

- Sessionization
- Recommendation
- Document Clustering

∞		
21h50m		First round of tuning
9h11m		Second round of tuning

memory space per CPU core impacts the maximum per-task heap space allocation that sessionization requires

http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Hadoop-Deployment-Comparison-Study.pdf

TUNING PARAMETERS

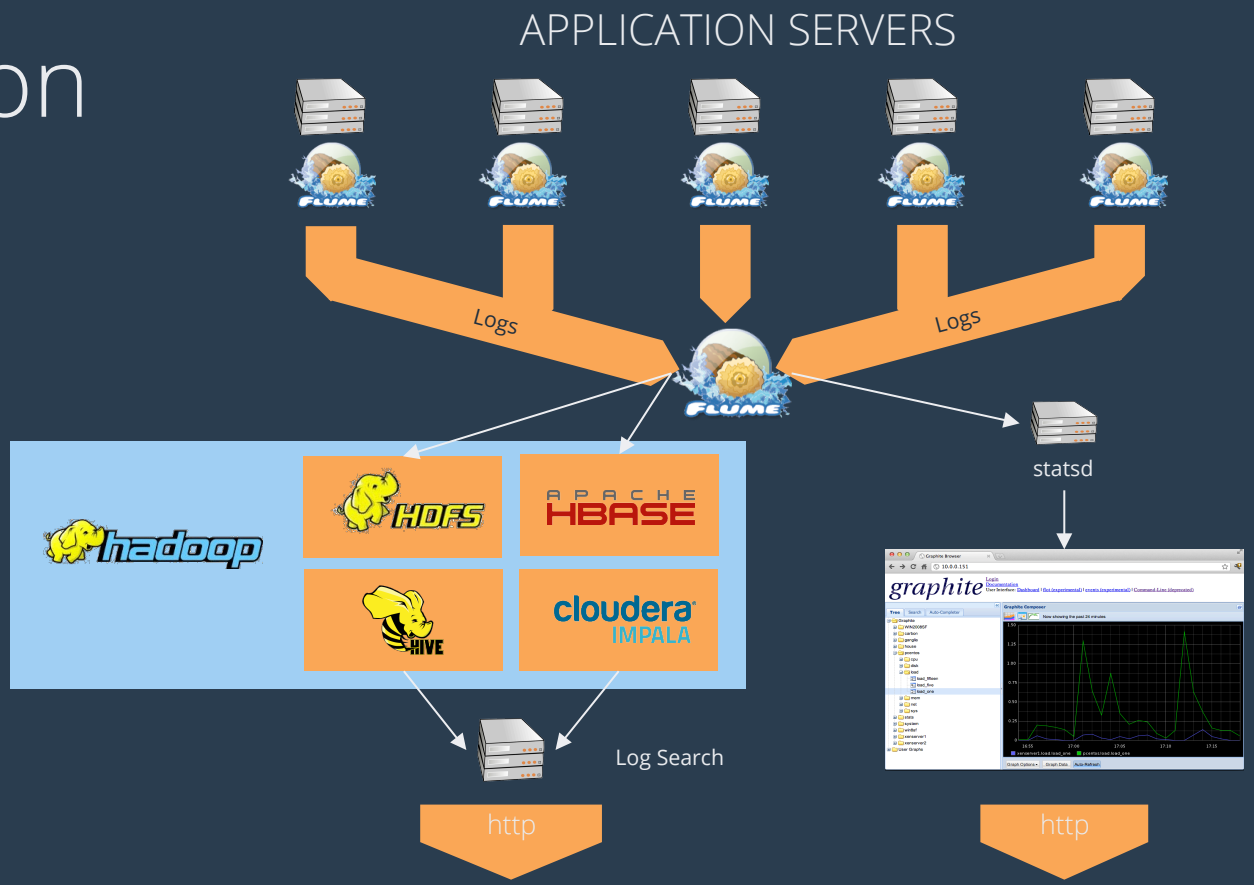
Cloud Specific

- Memory space per CPU core
- Number of CPU cores
- Storage per node

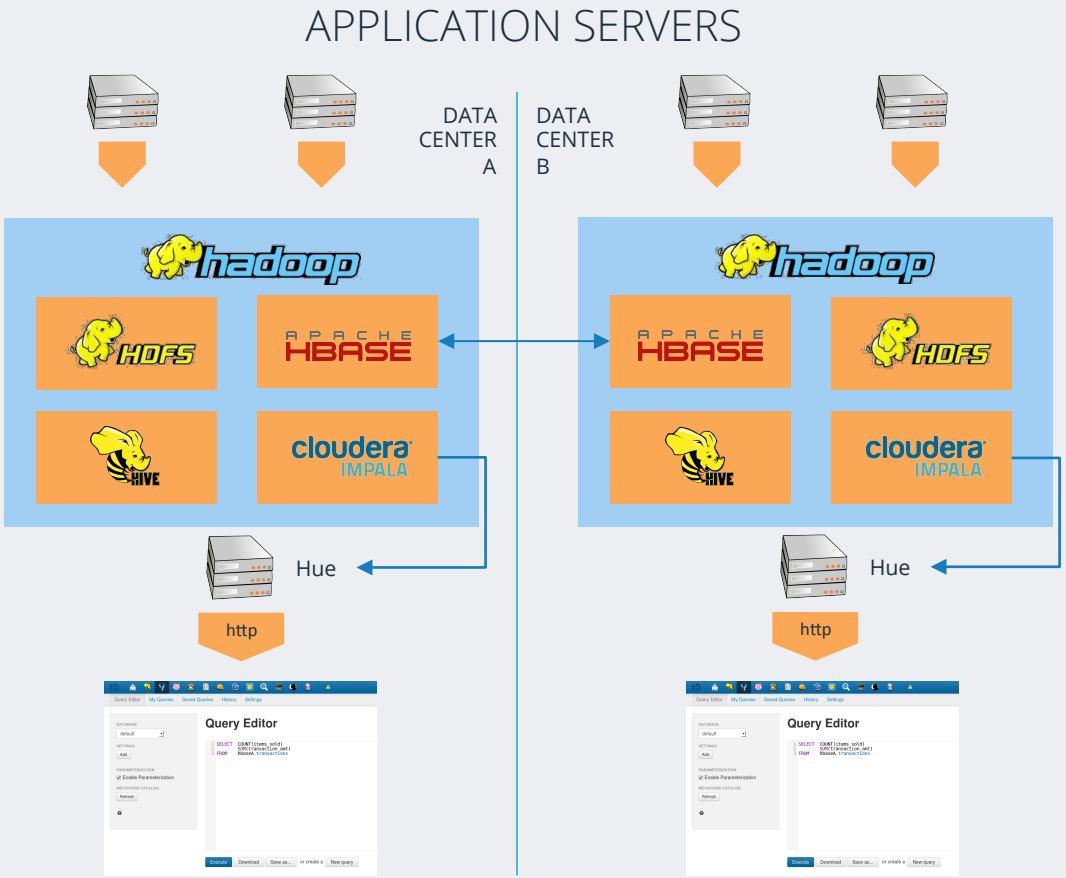
Cluster-Wide

- Number of map task slots and reduce task slots per TaskTracker node
- Task resource allocations

Log Collection & Search

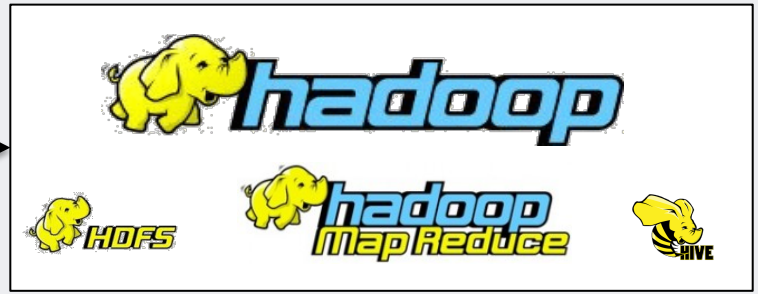


Real-Time Sales Transactions



Identity Matching

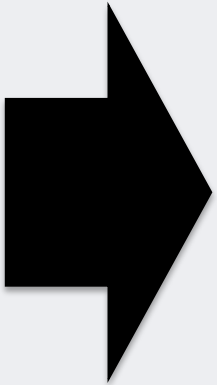
EDW

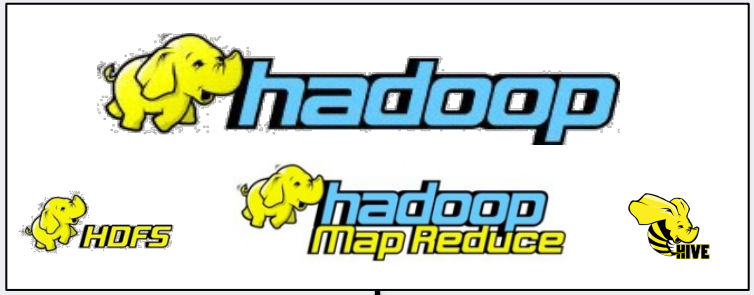


Push results to Cassandra

Sessionization

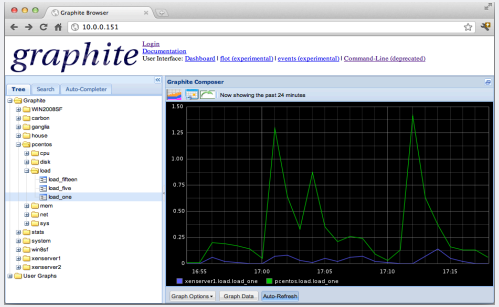
Clickstream





POLYGLOT PERSISTENCE

Horizontal
State

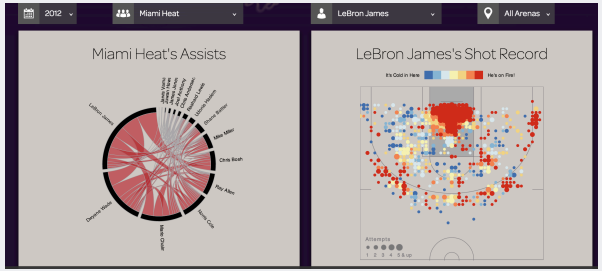


Operations
Supply Chain

Low Latency
Event Detection

	A139021	A123022	B5943	B5944	B5945
7/1/12 14:00	0.87	0.57	0.88	0.83	0.36
7/1/12 14:15	0.54	0.35			
7/1/12 14:30	0.16	0.72	0.68	0.77	0.77
7/1/12 14:45	0.02	0.74			
7/1/12 15:00	0.50	0.83	0.47	0.72	0.96
7/1/12 15:15	0.04	0.70			
7/1/12 15:30	1.00	0.31	0.22	0.62	0.93
7/1/12 15:45	0.07	0.29			
7/1/12 16:00	0.37	0.23	0.63	0.55	0.91
7/1/12 16:15	0.22	0.97			
7/1/12 16:30	0.69	0.65	0.02	0.23	0.91
7/1/12 16:45	0.53	0.96			
7/1/12 17:00	0.27	0.41	0.39	0.92	0.65
7/1/12 17:15	0.71	0.16			
7/1/12 17:30	0.67	0.35	0.34	0.41	0.67

Vertical
Historical

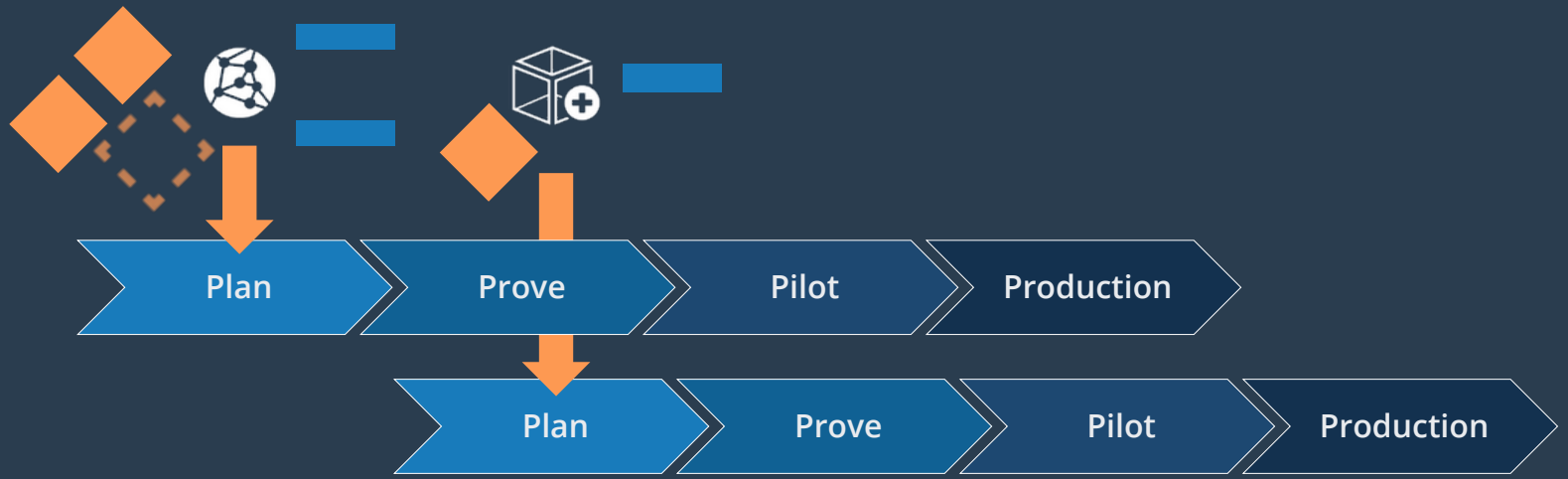


Planning
Asset Management

High Latency
Predictive Modeling

METHODOLOGY

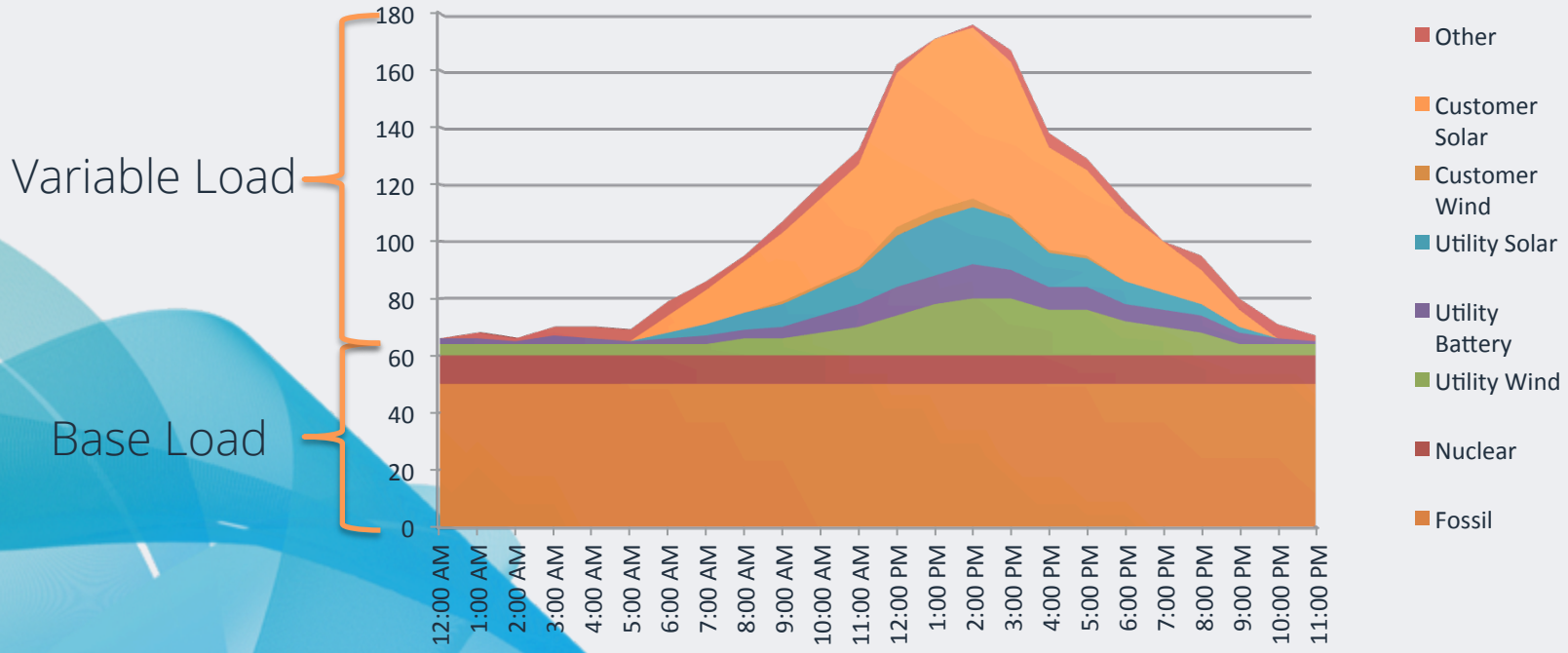
We iterate to value, answering the most valuable questions as quickly as possible



FROM EXPERIMENT TO DEPLOYMENT

- Pilot Workload
- Optimize Workload
- Analyze Production Requirements
- Determine TCO of Deployment Approaches
- Provision Deployment

Utility Source Mix



questions?

Yes, We're Hiring
svds.com

THANK YOU

John @BigDataAnalysis
Stephen @steveos

Slides are here: http://www.svds.com/downloads

SILICON VALLEY
DATA SCIENCE

O'REILLY®
Strata CONFERENCE + HADOOP WORLD

Tools and Techniques That Make Data Work